

Automatikus morfológiai elemzés a korai Alzheimer-kór felismerésében

Papp Petra Anna¹, Ráczi Anita¹, Vincze Veronika²

¹Szegedi Tudományegyetem, TTIK, Informatikai Tanszékcsoport,
Szeged, Árpád tér 2.

papp.petra.anna@gmail.com
raczanita89@gmail.com

²MTA-SZTE Mesterséges Intelligencia Kutatócsoport
vinczev@inf.u-szeged.hu

Kivonat: Cikkünkben Alzheimer-kórral diagnosztizált páciensek és egy egészséges kontrolcsoport beszédátíratának a magyarlanc által automatikus morfológiai elemzésnek alávetett változatait vizsgáljuk azzal a céllal, hogy a betegek nyelvhasználati sajátosságait felmérjük. Bemutatjuk a gépi elemzés kézi annotálással történő korrigálása során talált hibákat és javaslatokat teszünk a lehetséges javítási módokra. Emellett összefüggéseket keresünk a kórban szenvedők és egészséges társaik nyelvi produkciója között acélból, hogy egy, a jövőben megvalósuló rendszer keretén belül az esetlegesen érintettek beszédjük alapján felismerhetők legyenek.

Kulcsszavak: Alzheimer-kór, beszédátírat, morfológiai elemzés

1 Bevezetés

A szóbeli megnyilatkozások alapvetően különböznek az írásbeli formáktól, és olyan sajátosságokat is felmutatnak, amelyekre az írott szövegeken tanított nyelvi elemzők nincsenek kellőképpen felkészítve. Ez hatványozottan érvényes az Alzheimer-kórban szenvedő nyelvhasználók beszédére. A betegség beszédközpontra gyakorolt hatásának következtében nem csak a beteg számára okoz nehézséget a beszéd, hanem annak mások által történő megértése is akadályokba ütközik. A kórra utaló tünetek mihamarabbi észlelése annál is inkább lényeges, mert bár az Alzheimer-kór nem gyógyítható, mégis az időben érkező segítség nagyban javíthatja az érintettek életminőségét [1].

Ennek elősegítésére távlati célunk között szerepel egy olyan automatikus rendszer kifejlesztése, amely képes a korai Alzheimer-kórra jellemző nyelvi tünetek időben történő detektálására, így a beteg időben részesülhet a megfelelő kezelésben. Természetesen ennek a rendszernek nem célja, hogy helyettesítse az orvosi diagnózis felállítását, mégis fontos kiindulópontként szolgálhat annak megállapításában, hogy az Alzheimer-gyanús páciensek beszéde alapján ténylegesen fennáll-e a kór veszélye vagy sem. A rendszer beszélt nyelvi sajátosságokra, illetve a beszédátíratok automatikus morfológiai és szintaktikai elemzésén alapuló jellemzőkre épül.

Cikkünkben a beszédátíratok automatikus morfológiai elemzésére összpontosítunk. Vizsgálataink során azonosítottuk az elemző program számára nehézséget okozó

jelenségeket, melyek egyrészt a szóbeliség sajátosságaival magyarázhatóak, másrészt pedig magának az Alzheimer-kórnak tudhatóak be. A továbbiakban bemutatjuk a jellemző hibák csoportját, javaslatokat teszünk a lehetséges javítási megoldásokra. Emellett kitérünk arra is, hogy az egyes hibakategóriák előfordulási arányai alapján kimutatható-e szignifikáns különbség az Alzheimer-kóros betegek, illetve az egészséges kontrollcsoport nyelvhasználatá között.

2 A beszédátíratok

Az orvosok, beszédfeldolgozók, nyelvtechnológusok, informatikusok és pszichiáterek együttműködését kívánó feladat első lépését a nyelvi produkciók rögzítése képezte. Ennek kivitelezése orvosi felügyelet alatt történt az ún. Memóriaklinikán. A kutatás több hónapja folyamatosan zajlik az Alzheimer-kórral egyértelműen diagnosztizálható páciensek, valamint egészséges személyek körében. A vizsgáltak mindegyike azonos feladatot kapott: két rövid árnyjáték megtekintését követően össze kellett foglalniuk az első tartalmát, majd el kellett mesélniük a tegnapi napjukat. Utolsó feladatként a második videót összegezték. A páciensek teljes nyelvi produkciójáról hangfelvétel készült, melynek felhasználásához teljes beleegyezésüket adták. Az így létrejövő személyenkénti 3-3 monológot természetesen az elemzés teljes folyamán az adatvédelmi elveknek és személyiségi jogok védelmének megfelelően kezeltünk.

A következő lépésben az anonimizált hangfelvételek a Szegedi Tudományegyetem beszédfeldolgozóhoz kerültek, akik azokat elemezték és a beszédjellelmzők figyelembevételével, valamint azok kiejtэшű visszaadásával elkészítették szöveges átíratokat. Ebben a lépésben különleges hangsúlyt fektettek arra, hogy minden hangot és szót annak ténylegesen elhangzó formájában és hosszúságában adjanak vissza írásban, tehát a kiejtэшű sajátosságok mellett a beszédátíratok lényeges információkat hordoztak mind a beszédtempóról, mind a szünetekről, hezitációkról, nyújtott vagy téves szóindításokról stb.

Ezt követően a beszédátíratok a nyelvtechnológusokhoz kerültek, akik azokat több szempont szerint is elemezték. Egyfelől a beszéd során alkalmazott szókincset össze-sítettük és vizsgáltuk abból a célból, hogy a gyakori szavak lexikonjának összeállítá-sával még pontosabb képet nyerhessünk a kórral járó esetleges lexikai sajátosságok-ról. Ezt a kérdéskört a későbbiekben még részletesen tárgyaljuk.

Másrészt a beszédátíratokat automatikus morfológiai elemzésnek is alávetettük a magyarlanc szoftver [2] segítségével. A gépi elemzést kézi ellenőrzés követte, melynek során egyértelműen azonosítottuk azon problémacsoportokat, amelyek külön-ös nehézséget jelentettek az automatikus elemző számára. E jellemzők felismerése és kategóriákba sorolása azért oly lényeges, mert lehetővé teszi a magyarlanc felkészí-tését a kiejtэшű átíratok minél pontosabb elemzésére. Az azonosított hibák kategóriáit a következő pontban mutatjuk be.

3 Jellemző hibák a vizsgált személyek nyelvhasználatában

Ahogy arra már korábban utaltunk, a szóbeli megnyilatkozások jellemzői jelentős eltéréseket mutatnak az írásbeli formáktól. Ennek megfelelően a szóbeliség átirataiban bizonyos hibakategóriák összeállítására törekedtünk, melyeket a későbbiekben a morfológiai elemző finomításához kívánunk felhasználni. Jelen vizsgálatunk 27 Alzheimer-kórban szenvedő páciens, valamint 19 kontroll személy beszédátiratain alapul.

A vizsgált betegek jellemző tévesztései magukban foglalták a törléseket, hasonulásokat, a beszélők által létrehozott szavakat, a kettős szóindításokat, homofónokat, valamint hezitációkat. Az említett hibakategóriák nem mindegyike statisztikailag is szignifikáns indikátora azonban az Alzheimer-kórnak.

A törlések például az élőbeszéd tipikus megnyilvánulásai, és a betegek csoportjának beszédátirataiban is magasan reprezentáltak: *azé (azért), há (hát)*. A törlések bizonyos esetekben ugyanakkor homofóniához vezethetnek, azaz olyan alakokhoz, mint például *mer (mert)* vagy *mér (miért)*. Ez az eset azért tekinthető speciálisnak, mert e szavak ténylegesen nem homonimák, csupán a sztenderdtől eltérő kiejtésüknek és átiratuknak köszönhetően válnak kétértelművé, ami további nehézségeket jelent a megfelelő automatikus elemzés számára. Ezek a hibatípusok ugyanakkor nem tekinthetők a betegségre egyértelműen utaló tévesztéseknek, hiszen mind a sztenderd, mind a nyelvjárási beszélők nyelvhasználatának egyik jellemzője lehet.

Ugyanez érvényes a hasonulásokat alkalmazó nyelvhasználókra is, hiszen nem csupán az Alzheimer-kórosok, hanem az egészséges nyelvi beszélők is egyszerűsíthetik a szavakat ejtéskönnyítés céljából, így például természetes hangtani folyamatok következtében az *egyszer* gyakorta *eccer* alakban jelenik meg, míg a *képben* olykor *kébbe* alakot ölt.

Az új szavak alkotása a produktív nyelvhasználat egyik fontos jele, mely azonban szintén mindkét vizsgált csoportnál megfigyelhető volt. Az Alzheimer-kórral diagnosztizált betegek által kreált új kifejezések megfigyeléseink szerint azonban annyiban mégis különböznek az egészséges kontrollcsoportétól, hogy ezek jelentése kevésbé kikövetkeztethető a szövegkörnyezetből (pl. *sziriátum, bügyöre*). Mivel ez a kérdéskör további és tágabb vizsgálatokat igényel, ezért a morfológiai elemző finomítása egyelőre nem terjedt ki erre a problematikára.

Ugyanez igaz a kettős szóindításokra is, mint például *fé-férfi, asz-asztal*. Ezen alakok bár olykor az egészséges beszélőknél is fellelhetők, az Alzheimer-kóros pácienseknél különösen magas számban észleltük őket. Erre a jelenségre az elemző hibáinak manuális javításakor lettünk figyelmesek.

A betegek beszédében mégis a hezitációk száma és hosszúsága szolgálhat a legfontosabb tényezőként az egészséges társaiktól való megkülönböztetésben. Tény, hogy az élőbeszéd az egészséges személyek esetében is gyakorta hezitációkkal tüzelt, mégis a betegséggel diagnosztizáltak nyelvi produkciójában található hezitációk száma szignifikánsan különbözik a kontrollcsoport adataitól.

Erre vonatkozó adataink is ezt bizonyítják. A beszédíratok összesítéséből és elemzéséből kiderült, hogy az 549 hezitációs jel a kontrollcsoport esetében csaknem fele a beteg társaik által produkálténak (1005), a különbség statisztikailag szignifikáns ($p=0,0424$). Ezek alapján feltételezhető, hogy a hezitációk száma a kór fontos indikátora lehet és segítségünkre lehet annak felismerésében. Jelenlegi adataink szerint a diagnózis felállítása így nagyban épülhet a hezitációk gyakoriságára.

A következőkben példával is demonstráljunk az Alzheimer-kóros betegek és a normál beszélők megnyilvánulása közötti különbséget. A következő egy a kórral diagnosztizált beteg beszédátiratából származó részlet:

Baloldalt egy fie..fiatalember.. mászott le.. lepkehálóval.. amivel a lepkéket szokták megfogni. És .. avval átment a másik oldalra. És utána jött egy... nő... és a kosárba vót.. üveg.. poharak... és akkor ittak belőle egyegy valami volt benne amit ittak. Nem tudom mi lehetett... És abból ittak és utána elment mindakettő.

Ez alapján jól megfigyelhető a kettős szóindítások és a hezitációk gyakorisága is. A szöveg jellemző eleme az *és* kötőszó, illetve annak gyakori használata. A *valami* névmás és a *nem tudom* kifejezések itt is megjelennek, ezen egységek fontosságára még kitérünk. A szövegrészlet rövidebbé rámutat ugyanakkor arra, hogy komoly memóriazavarokkal küszködik a beszélő, hiszen a leglényegesebb szavakon kívül mást nemigen tudott visszaidézni.

A következő szövegrészlet egy egészséges személytől származik. Ebben is található néhány hezitálás, de lényegesen kevesebb, mint az előbbiben. Mint említettük, törlést az átlagos nyelvhasználók is elkövetnek, ezt jól reprezentálja a példa több pontja is.

Egy férfi és egy nő szerintem kirándulni mentek. És.. azt a röpködő valamit nem ismertem föl. Pillangónak kellett volna lenni, gondolom, mert olyan lepkéfogóval volt a férfi, éss azt szerette volna megfogni. Közbe egy ilyen mélyedésbe lekerült. Ott virágot tépett. Odaadta a szíve hölgyének, aki erre föl a kosarából egy flakon valamit, amit nem lehet tudni milyen itóka volt, odaadta. A férfi abból ivott utána a nő is ivott belőle. És... és ennyi.

4 Jellemző hibák a morfológiai elemzésben

Az írott szövegeken tanított automatikus morfológiai elemző komoly nehézségekbe ütközött a kiejtészű beszédátiratok elemzésekor. A gépi analízis kézi ellenőrzése során azonosított hibák alapvetően három, egymástól többé-kevésbé egyértelműen elkülöníthető kategóriába sorolhatóak. E csoportok szoros összefüggésben állnak az előző pontban említett nyelvi tévesztésekkel is, és egyértelmű elkülönítésük jelentős szerepet játszhat az elemző javításában.

Az első esetben X-es (ismeretlen) kóddal látja el az elemző azon megnyilatkozásokat, melyek morfológiai azonosításakor problémákba ütközik. Ez magában foglalja egyrészt a különböző hezitációkat jelölő formákat, másrészt pedig a törléseket, azaz a szóbeliséget és kiejtést hűen tükröző, és így a helyesírás szabályait felbontó átiratokat (pl. *há, azé*). A hezitációk magas aránya miatt külön figyelmet fordítottunk azon X-es kódú elemekre, amelyek hezitációra utalnak.

A hibaforrások másik nagy csoportját azon esetek alkotják, melyekben az elemző hibás (de létező) kódot rendel hozzá a szavakhoz. Kutatásaink szerint ez a jelenség a kettős szóindítások mellett főként az imént említett homofón szavaknál tapasztalható.

A harmadik osztályba végül a páciensek által újonnan képzett, és ily módon a nyelvhasználatban nem elterjedt egységeket soroltuk. Ezek esetében az elemző nem

jár el egységesen: többnyire ismeretlen szónak jelöli vagy a szóalak és toldalékai (pl. *szíriátum* latinra utaló *-um* végződése) alapján helyesen felismeri és elemzi azokat.

5 A magyarlanc átalakítási lehetőségei

A magyarlanc morfológiai és szintaktikai elemző a Szeged Korpusz [3] sztenderd nyelvhasználatú szövegein lett betanítva. A statisztikai elemzések alapján a beszédátiratokban talált ismeretlen szavak kb. 3,7%-os aránya jobban közelít a HunLearner [4] magyar nyelvtanulói korpuszban (szintén nem sztenderd magyar szövegekben) található ismeretlen szavak arányához (5,5%), mint a Szeged Korpuszéhoz (0,4%), amely számszerűsítve is alátámasztja a magyarlanc adaptációjának szükségességét a beszédátiratokra.

Az adaptációt elsődlegesen a morfológiai elemző szótárának kibővítésével kívánjuk megvalósítani, melyben elsődlegesen a beszédátiratokban talált hezitációkra fókuszálunk. Ez a folyamat tartalmazza azok kézi kigyűjtését a korpuszból és egyszerűsített átíratokat, melyben a hezitáció lehetséges megnyilvánulásait egységes alapokra hozzuk. Ezt a fajta normalizációt azért találtuk szükségesnek, mert az átíratokban a habozást jelölő alakok például a szókeresés időtartamára is utalnak, mint például *ööö* vagy *ööööööööö*. Célunk ezen alakok egy olyan prototipikus formával történő helyettesítése (pl. *öö*), mely segítségünkre lehet az eddig ismeretlenként kezelt egységek megfelelő felismerésében.

6 Eredmények

A jellemző hibák feltárásán és azok javítására vonatkozó javaslatainkon túl tanulmányunkban azt a célt is megfogalmaztuk, hogy különbségeket keressünk az Alzheimer-kórban szenvedő páciensek, valamint az egészséges kontrollcsoport között. Ehhez kapcsolódóan a következő táblázatban mutatjuk be a rendelkezésre álló beszédátiratok részleteit a vizsgált csoportok megoszlásában:

1. táblázat: A vizsgált csoportok beszédátiratainak összesített statisztikája

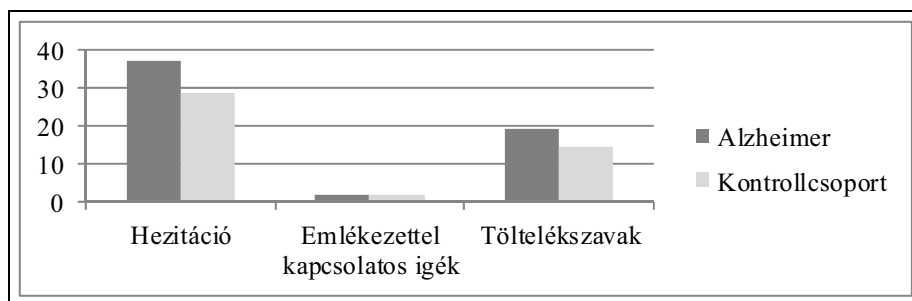
	Alzheimer	Kontroll
Mondatok száma	786	585
Átlagos mondat-szám egy beszélőre	29	31
Tokenek száma	11 614	8108
Átlagos tokenszám egy beszélőre	430	426

Az adatok egy szembetűnő tényről tűnnek alátámasztani. Az Alzheimer-kórral diagnosztizált betegek nyelvi produkciójának terjedelme ugyanis nem tér el lényegesen az egészséges kontrollcsoport által átlagosan használt mondat-, illetve szószámtól. Az

Alzheimer-kóros és az egészséges emberek megnyilatkozásai közötti különbség tehát kevésbé értelmezhető pusztán mennyiségi szempontból.

Az előbb közölt két beszédátírat-részletből is kitűnik azonban, hogy jelentős minősébeli eltérések jellemzik a két csoport átíratait.

Ahogy az a következő ábra is bemutatja, az Alzheimer-kóros páciensek beszédét gyakrabban kísérik töltelékszavak, a nyelvi redundancia és a hezitációk száma jelentősebb, mint a kontrollcsoportéban:



1. ábra: Alzheimer-kóros és egészséges nyelvhasználók hezitációinak, emlékezettel kapcsolatos megnyilatkozásainak és töltelékszavainak átlagos száma

Jelen statisztika három nyelvi jelenségről szolgáltat információt. Az első ilyen kategória a már korábban tárgyalt hezitációk és azok beszélőnkénti átlagos száma. A táblázatból kitűnik, hogy ez képezi a legjelentősebb minőségi különbséget a két vizsgált csoport megnyilatkozásaiban. Míg az egészséges emberek átlagosan körülbelül 29 alkalommal tartottak szünetet beszédükben, addig ez a szám az Alzheimer-kóros páciensek esetében ennél nyolc alkalommal, azaz csaknem 30 százalékkal gyakoribb volt. Ez az adat tehát egyértelműen alátámasztja azon korábban tett megállapításunk létjogosultságát, miszerint az Alzheimer-kór időben történő diagnózisában nagy szerepet játszhat a hezitációk gyakorisága.

A második, emlékezettel kapcsolatos megnyilatkozások csoportjába tartoznak az olyan kifejezések, mint például az *elfelejtettem* vagy a *nem tudom* (illetve ennek egyszerűsített alakja, a *nem tom*). Bár nem mutat szignifikáns különbséget, de valamivel jellemzőbb a betegekre, hogy a különböző kérdésekre *nem tudom* vagy *nem emlékszem* jellegű válaszokat adnak.

A töltelékszavak kategóriájába többek között a következő szavakat soroltuk: *ilyen, olyan, izé, és aztán, és akkor*, illetve a határozatlan névmásokat, úgymint *valamilyen, valahogy, valamerre*. Ezeket a szavakat nem bontottuk további csoportokba, mert ezek együttesen mutatnak rá, hogy az Alzheimer-kóros beszélők gyakran helyettesítésként szavakat határozatlan névmásokkal vagy az *izé* szócskával. Mellé knevek helyett pedig előszeretettel használnak parafrázisokat. Ennek megfelelően nem ritkák az *egy ilyen bagolyszerűség* vagy az *olyan délelőtt volt* körülíró, bizonytalanságra utaló kifejezések (vö. [5]), melyek jelenléte a harmadik pontban közölt Alzheimer-kóros beteg beszédében is megtalálhatóak. Az eredmények statisztikailag szignifikánsak ($p = 0,0316$), így egyfajta indikátora lehet a kórnak a töltelékszavak gyakori használata is.

7 Az eredmények felhasználása

Amint azt az előbbieken bemutattuk, a hezitációk, töltelékszavak előfordulási arányai szignifikáns eltéréseket mutatnak az Alzheimer-kórban szenvedők és az egészséges emberek esetében, így a fenti jellemzők jól hasznosíthatók a korai Alzheimer-kór diagnosztizálásában. Mindemellett ezek az eredmények a jövőben más nem sztenderd szövegek, például az internetes nyelvhasználat elemzésében is segítségünkre lehetnek.

A diagnózis felállítását megkönnyítendő olyan alapszókincs összeállításán dolgozunk, amely tartalmazza az egyes árnyjátékok tartalmának bemutatásakor használandó szavak listáját. Ebben az előforduló szavak gyakorisági statisztikáira, illetve a nyelvészek által összeállított listára építünk. Utóbbi szemléltetésére itt is közlünk egy részletet, amely azt mutatja be, hogy az egyik lejátszott filmhez kapcsolódó központi jelentőségű névszókkal általában milyen cselekvések társíthatóak a bemutatott kontextusnak megfelelően:

Névszó	Cselekvés
Férfi	<u>ül</u> ; <u>hív/int</u> a pincérnek; kettőt <u>mutat</u>
Pincér/felszolgáló	<u>Megérkezik/bejön/jön</u> ; <u>tisztítja/törli/söpri</u> az asztalt;
Gyűrű	Férfi <u>előveszi</u> , majd <u>kinyitja</u> <u>fényesíti</u> , <u>nézegeti</u>
Pincér	<u>Hoz / leteszi</u> a tálcán 2 poharat+1 üveget
Nő (esernyővel)	<u>Bejön/megérkezik</u> ; <u>kezet csókol</u> neki a férfi; nő <u>összecukja/becsukja</u> az ernyőjét, <u>leteszi</u> a székre
Pincér	<u>Kihúzza</u> a széket; <u>igazgatja</u> a széket
Férfi	<u>Int/szól</u> a pincérnek, hogy töltsön innivalót
Pincér	<u>Lenyomja</u> a nő fejét; nyakába <u>önti/leönti</u> a nőt
Férfi	<u>Felpattan</u> ; előrántja a botot/sétapálcát
Pincér	<u>Elbújik</u> a nő széke mögé; <u>vívna</u> k/kardoznak az esernyővel (pincér) és bottal (férfi)

Ez a lexikai oldalról való megközelítés azért is lesz központi jelentőségű, mert így lehetőségünk nyílik az Alzheimer-kórosok nyelvi produkciójának minőségbeli különbségeinek felmérésére.

Tervezzük emellett a rendelkezésre álló szótárunk további bővítését, és például további töltelékszavak gyűjtését, rendszerezését. A morfológiai hibák gyűjtésére és rendszerezésére is hangsúlyt szeretnénk a későbbiekben fektetni. Ennek keretein belül a szintaktikai hibákat is elemzés alá vonjuk a Szeged Dependency Treebank [6] vonatkereteinek segítségével.

Hosszabb távú terveink közé tartozik továbbá, hogy gépi tanulási eszközökkel a beszélőket beszédátírataik alapján elkülönítsük annak tekintetében, hogy fennáll-e náluk az Alzheimer-kór veszélye vagy sem.

8 Összegzés

Tanulmányunkban 27 Alzheimer-kórral diagnosztizált páciens és 19 egészséges személy nyelvi produkciójáról készült beszédátíratot vizsgáltunk acélból, hogy a nyelvi megnyilatkozásaikban alapvető különbségeket definiáljunk. Elemzéseink kiindulópontját a *magyarlanc* automatikus morfológiai elemző képezte, mely számára a szóbeli formák és különösen az alzheimeresek által produkált nyelvi alakok komoly kihívást jelentettek.

Ahogy az a statisztikák kimutatták, jelentős minőségbeli eltérések észlelhetők az egészséges és a beteg személyek nyelvi produkciójában, melyek főként a hezitációk és a töltelékszavak gyakoriságát foglalják magukban. Célunk, hogy ezen aspektusok figyelembevételével tovább finomítsuk az elemzőt. Meggyőződésünk, hogy ez nagyban hozzájárulhat ahhoz, hogy az esetlegesen érintettek esetében beszédük alapján mielőbb felismerhessük a kórt és így megfelelő kezelésben részesülhessenek.

Köszönetnyilvánítás

Jelen kutatást a Telemedicina fókuszú kutatások orvosi, matematikai és informatikai tudományterületeken című, TÁMOP-4.2.2.A-11/1/KONV-2012-0073 számú projekt támogatta. A projekt az Európai Unió támogatásával, az Európai Szociális Alap társfinanszírozásával valósul meg.

Hivatkozások

1. Hoffmann, I., Németh, D., Dye, Ch. D., Pákáski, M., Irinyi, T.; Kálmán, J.: Temporal parameters of spontaneous speech in Alzheimer's disease. In: *International Journal of Speech-Language Pathology* (2010) 12(1) 29–34
2. Zsibrita, J., Vincze, V., Farkas, R.: *magyarlanc: A Toolkit for Morphological and Dependency Parsing of Hungarian*. In: *Proceedings of RANLP-2013*, Hissar, Bulgaria (2013)
3. Csendes, D., Csirik, J., Gyimóthy, T., Kocsor, A.: The Szeged Treebank. In: *Proceedings of the Eighth International Conference on Text, Speech and Dialogue (TSD 2005)*. Karlovy Vary, Czech Republic 12–16 September, and LNAI series Vol. 3658 (2005) 123–131
4. Vincze V., Zsibrita J., Durst P., Szabó M. K.: HunLearner: a magyar nyelv nyelvtanulói korpusza. In: Tanács A., Vincze V. (szerk.): IX. Magyar Számítógépes Nyelvészeti Konferencia. Szegedi Tudományegyetem, Szeged (2013) 97–105

5. Vincze, V.: Bizonytalanságot jelölő kifejezések azonosítása magyar nyelvű szövegekben. In: Tanács A., Varga V., Vincze V. (szerk.): X. Magyar Számítógépes Nyelvészeti Konferencia. Szegedi Tudományegyetem, Szeged (2014)
6. Vincze, V., Szauter, D., Almási, A., Móra, Gy., Alexin, Z., Csirik, J.: Hungarian Dependency Treebank. In: Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC'10), Valletta, Málta (2010)